Statistica Sinica **7**(1997), 311-325

A CONSISTENT VARIABLE SELECTION CRITERION FOR LINEAR MODELS WITH HIGH-DIMENSIONAL COVARIATES

Xiaodong Zheng and Wei-Yin Loh

Utah State University and University of Wisconsin

Abstract: We consider the variable selection problem in regression models when the number of covariates is allowed to increase with the sample size. An approach of Zheng and Loh (1995) for the fixed design situation is extended to the case of random covariates. This yields a unified consistent selection criterion for both random and fixed covariates. By using t-statistics to order the covariates, the method requires much less computation than an all-subsets search. An application to autoregressive model selection with increasing order is given. The theory is supported by some simulation results.

Key words and phrases: Autoregressive processes, random covariates, t-statistic.

1. Introduction

Consider the linear regression model

$$y_i = x_{i,1}\beta_1 + \cdots + x_{i,M_n}\beta_{M_n} + \epsilon_i, \quad i = 1, \ldots, n \tag{1}$$

which relates a response variable y to a set of covariates $\{x_1, \ldots, x_{M_n}\}$. A frequently encountered problem is the selection of a subset of the covariates to keep in the final model. While it is important not to omit any true covariates (i.e., those whose population regression coefficients are nonzero), it is well known that inclusion of non-true or nuisance covariates generally reduces model prediction accuracy. Another consequence of selecting only the true covariates is that the complexity of the final regression model is reduced. This is especially desirable when the total number M_n of available covariates is large. A good general discussion of the effect of variable selection on parameter estimation and prediction is given in Miller (1990). Let

$$\Gamma = \{i \mid \beta_i \neq 0, 1 \leq i \leq M_n\} \tag{2}$$

be the index set of the true covariates (also called the true model). Then the variable selection problem is equivalent to estimating Γ, which in this article is assumed to be independent of n.

When the covariates are assumed to be nonrandom, many selection criteria have been proposed and studied. These include the FPE criterion (Thompson (1978), Shibata (1984), Zhang (1992)), cross-validation (Burman (1989), Zhang (1993), Shao (1993)) and bootstrap methods (Efron (1983), Zheng and Loh (1994)). Although the FPE criterion is inconsistent, it has the advantage (compared to the resampling methods) of being easy to use and fast to compute. A class of consistent criteria that retain the simple form of the FPE is developed in Zheng and Loh (1995).

In this paper, we consider the variable selection problem when the covariates are random. Random covariates arise in many regression applications where the values of the covariates can only be observed and are not controllable. The importance of variable selection in this case is clearly recognized by Breiman and Spector (1992), who argue that models with random covariates typically have substantially higher prediction errors than the fixed design counterparts and hence more is gained by variable selection. Other distinctions between the two different models can be found in Thompson (1978).

The traditionally recommended criterion for random design models is the S_p method (Hocking (1976), Thompson (1978), Linhart and Zucchini (1986)). Let Θ be the set of the indices of variables in a particular model. Then the S_p method selects the model $\hat{\Gamma}_{S_p}$ which minimizes

$$S_p(\Theta) = (n - |\Theta|)^{-1}(n - |\Theta| - 2)^{-1}\mathrm{RSS}(\Theta)$$

over all submodels $\Theta \subseteq \Omega = \{1, \ldots, M_n\}$. Here $|\Theta|$ is the cardinality of Θ and $\mathrm{RSS}(\Theta)$ is the residual sum of squares for model Θ fitted by the least squares method. The justification for the S_p criterion is that, under joint normality of covariates and the regression error, $n^{-1}(n-2)(n+1)S_p(\Theta)$ is an unbiased estimator of the expected square prediction error for model Θ. In spite of this, the estimator $\hat{\Gamma}_{S_p}$ is generally not consistent for the true model Γ in the sense that

$$\lim_{n\to\infty} P(\hat{\Gamma}_{S_p} = \Gamma) \neq 1 \tag{3}$$

(see Breiman and Freedman (1983) for a different conclusion under the setting when the true model contains infinitely many covariates). Although certain statisticians are aware of this deficiency of the S_p criterion, a rigorous proof has not been given in the literature. We give a sketch of it in the Appendix. The proof also shows that the S_p criterion is able to eliminate underfitting but not overfitting models. A similar conclusion can be drawn for the FPE criterion, which estimates the true model by minimizing

$$\mathrm{FPE}(\Theta) = \mathrm{RSS}(\Theta) + \lambda|\Theta|\hat{\sigma}^2, \;\; \Theta \subseteq \Omega, \tag{4}$$

where λ is a positive constant and

$$\hat{\sigma}^2 = (n - M_n)^{-1}\mathrm{RSS}(\Omega) \tag{5}$$

is the usual estimate of the error variance σ^2 based on the full model.

To search for a viable solution, we study the class of criteria proposed in Zheng and Loh (1995) and show that they remain consistent. This provides a unified consistent variable selection approach to both the fixed and the random design situations. Our method is developed in Section 2, where we first consider the simple situation when the covariates are pre-ordered such that the true covariates are indexed before the nuisance covariates. It is then generalized to the unordered case by use of t-statistics to order the covariates. In both instances, we allow M_n to grow with the sample size n. This flexibility is important since in many applications the number of covariates is usually not small relative to n; (see e.g., Huber (1981) and Bickel and Freedman (1983)).

We apply our method to estimation of the true dimension of an autoregressive (AR) process in Section 3. Our results permit the selection process to include increasing-dimensional AR models. The latter is useful in practice because the true dimension is often unknown. Some simulation results to support the asymptotic theory are reported in the last section, followed by an appendix containing all the technical details.

2. A Class of Consistent Criteria

The inconsistency of the S_p and FPE criteria is not unusual. A similar phenomenon also exists for the FPE selection criterion in fixed design regression (Zhang 1992). As argued by Zheng and Loh (1995), the major source of the FPE's deficiency lies in the insufficient amount of penalty it places on $\lambda|\Theta|\hat{\sigma}^2$ for model complexity. They show that replacement of $\lambda|\Theta|\hat{\sigma}^2$ in (4) by a penalty term of the form $h_n(|\Theta|)\hat{\sigma}^2$ leads to a consistent selection criterion.

We now extend this approach to regression model (1) with high dimensional random covariates. Although modification of the S_p criterion could be an alternative, the current approach has the advantage of providing a unified variable selection criterion that is consistent for both random and fixed design cases. Another feature of our method is that, by using the regression t-statistics to order the covariates, one only needs to search M_n instead of all 2^{M_n} subsets of covariates for the true model. The computational savings are substantial when there are large numbers of covariates.

To fix some additional notation, rewrite model (1) as $y_i = \mathbf{x}_i'\beta + \epsilon_i$, $i = 1, \ldots, n$, where $\mathbf{x}_i' = (x_{i,1}, \ldots, x_{i,M_n})$ and $\beta' = (\beta_1, \ldots, \beta_{M_n})$, the dependence on n being understood. For any submodel Θ let β_Θ be the sub-vector of β with components β_k, $k \in \Theta$. Similarly, we let $\mathbf{x}_{i,\Theta}$ and $\mathbf{X}_\Theta = (\mathbf{x}_{1,\Theta}, \ldots, \mathbf{x}_{n,\Theta})'$

denote the corresponding ith design sub-vector and design matrix respectively. Note that the projection matrix $\mathbf{P}_\Theta = \mathbf{X}_\Theta(\mathbf{X}'_\Theta\mathbf{X}_\Theta)^-\mathbf{X}'_\Theta$ is invariant for any generalized inverse $(\mathbf{X}'_\Theta\mathbf{X}_\Theta)^-$. Throughout, we drop the subscript Θ whenever $\Theta = \Omega$, the full model.

The following minimum assumptions will be in effect in this section.

(A1) $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are i.i.d. with finite second moment and are independent of $\epsilon = (\epsilon_1, \ldots, \epsilon_n)'$, which are i.i.d. with mean 0 and variance σ^2.

(A2) $E(\mathbf{x}_1\mathbf{x}'_1)$ is nonsingular. (This implies that $E(\mathbf{x}_{1,\Theta}\mathbf{x}'_{1,\Theta})$ is nonsingular for every Θ.)

Condition (A2) is necessary to ensure that β uniquely minimizes the expected square error function $f(\mathbf{b}) = E(y_i - \mathbf{x}'_i\mathbf{b})^2$.

2.1. The pre-ordered case

Consider first the special case when the covariates are pre-ordered such that the true covariates are indexed before the nuisance covariates. That is, the true β in (1) takes the form $\beta' = (\beta_1, \ldots, \beta_{k_0}, 0, \ldots, 0)$ for some k_0 independent of n and $\beta_i \neq 0$, $i \leq k_0$. Then searching for the true model Γ defined by (2) is equivalent to estimating the unknown k_0. For ease of notation, we shall denote $\Theta_k = \{1, \ldots, k\}$ and $\Theta_0 = \emptyset$. Note that $|\Theta_k| = k$. We also write RSS_k for $\mathrm{RSS}(\Theta_k)$ and $\mathbf{P}_k$ for $\mathbf{P}_{\Theta_k}$.

Define

$$\hat{k}_0 = \arg\min_{0 \leq k \leq M_n} \{\mathrm{RSS}_k + h_n(k)\hat{\sigma}^2\}, \tag{6}$$

where $\hat{\sigma}^2$ is defined in (5) and $h_n(k)$ is a nonnegative function.

We impose the following conditions on h_n and M_n.

(B1) $M_n/n \to 0$ as $n \to \infty$.

(B2) $h_n(k)$ is nondecreasing in k with $h_n(0) = 0$ and $\liminf_n h_n(k+1)/h_n(k) > 1$ for any $k \geq 1$.

(B3) For each $k \geq 1$, $h_n(k)/M_n \to \infty$ as $n \to \infty$.

(B4) For each $k \geq 1$, $n^{-1}h_n(k) \to 0$ as $n \to \infty$.

Theorem 1. *Under conditions* (A1), (A2) *and* (B1)–(B4), $\hat{k}_0 \to_P k_0$ *as* $n \to \infty$.

Remark.

1. The two most important conditions here are (B3) and (B4), which act in opposite directions. The appropriate growth rate of $h_n(k)$ is between M_n and n. Thus the number of covariates plays a critical role, in the sense that a heavier penalty function h_n is required when there are many nuisance covariates. Note also that the existence of h_n satisfying (B3) and (B4) is guaranteed by (B1).

2. As a direct application of Theorem 1, we have that the BIC criterion (with $h_n(k) = k \log n$) is consistent if $M_n = o(\log n)$. Furthermore, if $M_n = o(\log \log n)$ then setting $h_n(k) = k \log \log n$ in (6) leads to consistency of the ϕ criterion of Hannan and Quinn (1979). On the other hand, if M_n is large compared to n, the performance of these two criteria can be inadequate and a heavier penalty is called for; see the simulation results in Section 4.
3. Because M_n may increase with n, the assumptions of Theorem 1 are not sufficient for the asymptotic existence of $(\mathbf{X}'\mathbf{X})^{-1}$ unless some additional moment conditions are imposed (see Section 2.2). However, it is always true that $\mathrm{tr}(\mathbf{P}) = \mathrm{rank}(\mathbf{X}) \leq M_n$ a.s. On the other hand, since $\mathbf{X}'_{k_0}\mathbf{X}_{k_0}$ has fixed dimension k_0, it is of full rank for large n a.s. Thus Theorem 1 allows for the presence of collinear nuisance covariates.

2.2. The general case

For the general case when the regression covariates are not pre-ordered, a natural approach is to first order them consistently and then apply criterion (6). An appealing property of this approach is that it involves much less computation than an all-subsets search.

Let $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_{M_n})'$ be the least squares estimators based on the full model and let

$$T_i = \hat{\sigma}^{-1}\hat{\beta}_i\{i\text{th diagonal element of } (\mathbf{X}'\mathbf{X})^{-1}\}^{-1/2}, \quad i = 1, \ldots, M_n,$$

be the corresponding t-statistics. We assume the following conditions, which imply that with probability tending to one, $(\mathbf{X}'\mathbf{X})^{-1}$ exists and therefore $\hat{\beta}$ is unique. Similar conditions are used by Mammen (1993) to study the asymptotic behavior of bootstrap linear contrasts in increasing-dimension linear models.

(C1) The minimum eigenvalue κ_n of $E(\mathbf{x}_1\mathbf{x}_1')$ is bounded away from zero, i.e., $\kappa_n \geq \kappa > 0$ for some $\kappa > 0$.

(C2) For some $\eta > 0$, $n^{-1}M_n^{1+\eta} \to 0$ and

$$\sup_n \sup_{\|d\|=1} E|d'[E(\mathbf{x}_1\mathbf{x}_1')]^{-1/2}\mathbf{x}_1|^{4[2/\eta]} < \infty, \tag{7}$$

where $[2/\eta]$ is the smallest integer greater than or equal to $2/\eta$.

Note that the smaller M_n is, the weaker the moment condition (7) becomes. Further, if $\mathbf{x}_1$ has a zero mean normal distribution as assumed in Thompson (1978), then $d'[E(\mathbf{x}_1\mathbf{x}_1')]^{-1/2}\mathbf{x}_1$ is a standard normal random variable and η can be chosen to be arbitrarily small.

The proposed variable selection procedure goes as follows:

Step 1. Sort the covariates in order of decreasing absolute values of the t-statistics:

$$|T_{i_1}| \geq |T_{i_2}| \geq \cdots \geq |T_{i_{M_n}}|.$$

Step 2. Apply criterion (6) to the ordered covariates $i_1, i_2, \ldots, i_{M_n}$. That is, estimate the true model Γ by

$$\hat{\Gamma} = \{i_1, \ldots, i_{\hat{k}^*}\}, \tag{8}$$

where $\hat{k}^* = \arg\min_{0 \leq k \leq M_n} \{\mathrm{RSS}^*(k) + h_n(k)\hat{\sigma}^2\}$ and $\mathrm{RSS}^*(k)$ is the residual sum of squares for model $\{i_1, \ldots, i_k\}$.

The ordering procedure using t-statistics was proposed, under the name "$t_{K,i}$-directed search", in Daniel and Wood (1980), Chapter 6 as a possible tool for reducing computation in stepwise regression. However, no theoretical justification for consistency was given there. For a discussion of model selection after ordering of covariates by t-tests, see An and Gu (1985). The following theorem is a generalization of a theorem of Zheng and Loh (1995) from the fixed design case with a fixed number of covariates to the random design situation with the number of covariates allowed to depend on n.

Theorem 2. *Suppose conditions* (C1)–(C2) *and the assumptions of Theorem* 1 *hold. Then*

$$\lim_{n\to\infty} P(\min_{i\in\Gamma} |T_i| > \max_{i\notin\Gamma} |T_i|) = 1 \tag{9}$$

and criterion (8) *is consistent for* Γ, *i.e.,* $\lim_n P(\hat{\Gamma} = \Gamma) = 1$.

3. Autoregressive Model Selection

The proposed criterion in Section 2.1 can be applied to model selection in time series. We shall discuss this application in the framework of autoregressive processes.

Let $\{y_i,\ 1-M_n \leq i \leq n\}$ be an autoregressive process of order M_n ($\mathrm{AR}(M_n)$) satisfying

$$y_i = \beta_1 y_{i-1} + \cdots + \beta_{M_n} y_{i-M_n} + \epsilon_i,\ i = 1, \ldots, n, \tag{10}$$

where ϵ_i are i.i.d. with mean zero and variance σ^2. Assume that the true model is $\mathrm{AR}(k_0)$, that is, $\beta_{k_0} \neq 0$, $\beta_j = 0$, $j > k_0$ and k_0 is independent of n.

The literature on the subject of estimating k_0 is extensive. See, for example, Akaike (1974), Hannan and Quinn (1979), Hannan (1980), Rissanen (1986) and Wei (1992). Choi (1992) gives a comprehensive survey. All existing results require a good guess of an upper bound $M \geq k_0$ and a search for the true model over $\mathrm{AR}(p)$, $0 \leq p \leq M$. They also treat M as fixed. This formulation is not very practical, however. Since k_0 is unknown, it is conceivable that the upper bound should be larger than a small fraction of n. To guarantee that $M \geq k_0$, it is therefore necessary and more natural to consider the problem in such a way that $M = M_n$ may increase with n. We give a consistent theory for this case here. The result depends on the study of increasing-dimensional AR models.

Rewrite model (10) as $y_i = \mathbf{x}_i'\beta + \epsilon_i$, where $\mathbf{x}_i' = (y_{i-1}, \ldots, y_{i-M_n})$. Then all the previous notation applies. For instance, $\hat{k}_0$ is defined in (6), where RSS_k is the residual sum of squares of the AR(k) model fitted by the least-squares method. To avoid some technicalities, we also assume that the process $\{y_i\}$ is stationary and normally distributed with zero mean. Stationarity implies that $\mathbf{x}_1, \ldots, \mathbf{x}_n$ have the same distribution, that the characteristic equation in z

$$h(z) = 1 - \sum_{j=1}^{k_0} \beta_j z^j = 0 \tag{11}$$

has all k_0 roots outside the unit circle in the complex plane, and that the autocorrelation $\rho_{k-j} = \rho_{j-k}$ between y_k and y_j satisfies

$$|\rho_{k-j}| \leq C a^{|k-j|} \tag{12}$$

for some constants $C > 0$ and $0 < a < 1$ depending only on k_0 and β_{k_0} (Box and Jenkins (1976), Section 3.2). In addition, we require that
(B1') $M_n^2/n \to 0$.
This condition ensures that the $M_n \times M_n$ matrix $n^{-1}\mathbf{X}'\mathbf{X}$ is a good estimate of its expectation (c.f. Lemma 1 below). It may be possible to weaken this to a condition like (C2) in Section 2.2 using graph theory (Mammen (1993), Lemma 1). Such an improvement is quite minor however and in practice a moderately large value M_n satisfying (B1') should provide a satisfactory upper bound on the true model dimension.

Theorem 3. *Under conditions* (B1') *and* (B2)–(B4) *in Section* 2.1, $\hat{k}_0 \to_P k_0$.

4. A Simulation Study

To compare the finite-sample performance of the variable selection procedures, we carried out a simulation experiment with 1,000 trials and sample size $n = 300$ per trial.

For ordinary regression, we used the models (i) $\Gamma = \{2, 4, 5\}$, (ii) $\Gamma = \{k^2; k = 1, \ldots, 5\}$ and (iii) $\Gamma = \{2k + 7; k = 3, \ldots, 12\}$, with $M_n = 5$, 30 and 60, respectively. Each β_k, $k \in \Gamma$, was set equal to 1. The covariates were generated by a M_n-variate zero mean normal distribution with the (i, j)th entry of the covariance matrix being $2^{-|i-j|}$. The distribution of ϵ_i was standard normal. For criterion (8), we used $h_n(k) = kn^{0.3}$ for $M_n = 5$ and $h_n(k) = kn^{0.7}$ for $M_n = 30$ and 60. For the latter two values of M_n, the covariates were pre-ordered by their t-statistics for all four selection criteria because an all-subsets search was impractical. The simulation was coded in FORTRAN using a singular value decomposition subroutine and carried out on a SUN SPARCstation 20.

For AR model selection, we tested two cases: $M_n = 5$ and 30. For $M_n = 5$, the true model was AR(1): $y_i = -0.3y_{i-1} + \epsilon_i$, and $h_n(k) = kn^{0.3}$. For $M_n = 30$, the true model was AR(10): $y_i = 0.2y_{i-10} + \epsilon_i$, and $h_n(k) = kn^{0.7}$. For both cases, ϵ_i were i.i.d. N(0,1). Initial values $y_t, t = 1 - M_n, \ldots, 0$, were set to zero.

Table 1. Estimated probabilities of correct model selection based on 1,000 trials; $n = 300$. The proposed criterion (8) is given in the last column of the table.

M_n	True model Γ	S_p	AIC	BIC	Proposed
5	$\{2, 4, 5\}$	0.533	0.502	0.916	0.971
30	$\{k^2, k = 1, \ldots, 5\}$	0.377	0.365	0.754	0.929
60	$\{2k + 7, k = 3, \ldots, 12\}$	0.159	0.173	0.362	0.892
Est. max. standard error		0.016	0.016	0.015	0.010

Table 1 summarizes the results for ordinary regression, which include those for the S_p, the AIC (i.e., the FPE with $\lambda = 2$) and the BIC criteria. The S_p and AIC criteria have fairly low probabilities of correct model selection. The BIC criterion performs better for small and moderate values of M_n, but it is poor when there are many nuisance covariates ($M_n = 60$). This shows the necessity for placing a heavier penalty on model complexity. The best procedure is clearly our proposed criterion (8).

Table 2. Estimated probabilities of correct AR model selection based on 1,000 trials; the proposed criterion (6) is given in the last column of the table.

M_n	AIC	BIC	Proposed
5	0.554	0.898	0.956
30	0.296	0.790	0.914
Est. max. S.E.	0.016	0.012	0.009

Table 2 gives the corresponding results for AR model selection. Again the proposed method is best.

Acknowledgement

This research was supported in part by U. S. Army Research Office grant DAAH04-94-G-0042.

Appendix

Proof of (3).

We only give an outline. The notation defined in Section 2 is used. Assume conditions (A1) and (A2) and that $M_n = M$ is independent of n. Then asymptotic expansions give

$$S_p(\Theta) = \begin{cases} n^{-2}\epsilon'\epsilon + n^{-2}\{2\sigma^2(|\Theta| + 2) - \epsilon'\mathbf{P}_\Theta\epsilon\} + o_P(n^{-2}), & \Theta \supseteq \Gamma \\ n^{-2}\epsilon'\epsilon + n^{-1}\beta'\mathbf{\Sigma}_\Theta\beta + o_P(n^{-1}), & \Theta \not\supseteq \Gamma. \end{cases} \tag{13}$$

Here $\mathbf{\Sigma}_\Theta = \mathbf{C}_\Theta \mathrm{diag}(\mathbf{0}, \mathbf{\Psi}_\Theta)\mathbf{C}'_\Theta$, where

$$\mathbf{\Psi}_\Theta = E(\mathbf{x}_{1,\Theta^c}\mathbf{x}'_{1,\Theta^c}) - E(\mathbf{x}_{1,\Theta^c}\mathbf{x}'_{1,\Theta})[E(\mathbf{x}_{1,\Theta}\mathbf{x}'_{1,\Theta})]^{-1}E(\mathbf{x}_{1,\Theta}\mathbf{x}'_{1,\Theta^c})$$

is positive definite (Seber (1984), Exercise 2.20), Θ^c is the complement of Θ, and $\mathbf{C}_\Theta$ is a permutation matrix such that $\mathbf{x}_1 = \mathbf{C}_\Theta(\mathbf{x}'_{1,\Theta}, \mathbf{x}'_{1,\Theta^c})'$ and $\mathbf{C}_\Theta^{-1} = \mathbf{C}'_\Theta$ (see, e.g., Golub and Van Loan (1989), Section 3.4.1).

For each $\Theta \not\supseteq \Gamma$, we have $\beta'\mathbf{\Sigma}_\Theta\beta = \beta'_{\Theta^c}\mathbf{\Psi}_\Theta\beta_{\Theta^c} > 0$ and $\epsilon'\mathbf{P}_\Theta\epsilon = O_P(1)$. Consequently, $\lim_n P(\hat{\Gamma}_{S_p} = \Theta) = 0$ for every $\Theta \not\supseteq \Gamma$, i.e., the S_p criterion eliminates underfitting models $\Theta \not\supseteq \Gamma$.

On the other hand, if the true model is not the full model one can choose a $\Theta \supset \Gamma$. Suppose additionally that the ϵ_i are normally distributed. Then

$$\begin{aligned}\lim_n P(\hat{\Gamma}_{S_p} = \Gamma) &\le \lim_n P[\sigma^{-2}\epsilon'(\mathbf{P}_\Theta - \mathbf{P}_\Gamma)\epsilon \le 2(|\Theta| - |\Gamma|)] \\ &= P[\chi^2_{|\Theta|-|\Gamma|} \le 2(|\Theta| - |\Gamma|)] < 1.\end{aligned}$$

This proves the inconsistency of the S_p criterion when M_n does not depend on n and the ϵ_i are normal. The general case when M_n may grow with n follows similarly.

Proof of Theorem 1.

Define $\tilde{k}_0 = \arg\min_{k_0 \le k \le M_n}\{\mathrm{RSS}_k + h_n(k)\hat{\sigma}^2\}$. The proof proceeds in two steps.

1. $\tilde{k}_0 \to_P k_0$. For $k \ge k_0$, the residual sum of squares reduces to $\mathrm{RSS}_k = \epsilon'\epsilon - \epsilon'\mathbf{P}_k\epsilon$. Condition (B1) implies that $E\{\epsilon'\mathbf{P}\epsilon/(n - M_n)\} \le \sigma^2 M_n/(n - M_n) = o(1)$ and hence

$$\hat{\sigma}^2 = \epsilon'\epsilon n^{-1}(1 - M_n/n)^{-1} - \epsilon'\mathbf{P}\epsilon/(n - M_n) = \sigma^2 + o_P(1). \tag{14}$$

By condition (B2) and the fact that $(\mathbf{P}_{M_n} - \mathbf{P}_k)$ is idempotent a.s., we have for $k > k_0$,

$$\begin{aligned}&\mathrm{RSS}_k + h_n(k)\hat{\sigma}^2 - \mathrm{RSS}_{k_0} - h_n(k_0)\hat{\sigma}^2 \\ &\ge \hat{\sigma}^2[h_n(k_0 + 1) - h_n(k_0)] - \epsilon'(\mathbf{P}_{M_n} - \mathbf{P}_{k_0})\epsilon.\end{aligned}$$

Therefore,

$$
\begin{aligned}
&1 - P(\tilde{k}_0 = k_0) \le 1 - P\{\min_{k>k_0}[\mathrm{RSS}_k + h_n(k)\hat{\sigma}^2 - \mathrm{RSS}_{k_0} - h_n(k_0)\hat{\sigma}^2] > 0\} \\
&\le P[\epsilon'(\mathbf{P}_{M_n} - \mathbf{P}_{k_0})\epsilon \ge \hat{\sigma}^2\{h_n(k_0+1) - h_n(k_0)\}] \\
&\le P[\epsilon'(\mathbf{P}_{M_n} - \mathbf{P}_{k_0})\epsilon > \{h_n(k_0+1) - h_n(k_0)\}\sigma^2/2] + P(|\hat{\sigma}^2 - \sigma^2| \ge \sigma^2/2) \\
&\le 2[h_n(k_0+1) - h_n(k_0)]^{-1}\sigma^{-2}E\{\epsilon'(\mathbf{P}_{M_n} - \mathbf{P}_{k_0})\epsilon\} + o(1) \qquad (15) \\
&\le 2[1 - h_n(k_0)/h_n(k_0+1)]^{-1}[M_n/h_n(k_0+1)] + o(1) \\
&\to 0, \qquad (16)
\end{aligned}
$$

where (15) follows from the Markov inequality and (14), and (16) is a consequence of conditions (B2) and (B3).

2. $\hat{k}_0 - \tilde{k}_0 = o_P(1)$. Using $\mathrm{RSS}_j = \epsilon'(\mathbf{I} - \mathbf{P}_j)\epsilon + \beta'\mathbf{X}'(\mathbf{I} - \mathbf{P}_j)\mathbf{X}\beta + 2\epsilon'(\mathbf{I} - \mathbf{P}_j)\mathbf{X}\beta$ and $\mathbf{P}_{k_0}\mathbf{X}\beta = \mathbf{X}_{k_0}\beta_{k_0} = \mathbf{X}\beta$ a.s., we get

$$
\begin{aligned}
&P(|\hat{k}_0 - \tilde{k}_0| \ne 0) \le \sum_{j=0}^{k_0-1} P(\hat{k}_0 = j) \\
&\le \sum_{j=0}^{k_0-1} P\{\mathrm{RSS}_j + h_n(j)\hat{\sigma}^2 \le \mathrm{RSS}_{k_0} + h_n(k_0)\hat{\sigma}^2\} \\
&\le \sum_{j=0}^{k_0-1} P\{-2\epsilon'(\mathbf{I} - \mathbf{P}_j)\mathbf{X}\beta \ge \|(\mathbf{I} - \mathbf{P}_j)\mathbf{X}\beta\|^2 - h_n(k_0)\hat{\sigma}^2\} \\
&\le \sum_{j=0}^{k_0-1} P\{-2\epsilon'(\mathbf{I} - \mathbf{P}_j)\mathbf{X}\beta \\
&\ge \|(\mathbf{I} - \mathbf{P}_j)\mathbf{X}\beta\|^2 - h_n(k_0)(3/2)\sigma^2\} + k_0 P(|\hat{\sigma}^2 - \sigma^2| > \sigma^2/2) \\
&= \sum_{j=0}^{k_0-1} P\{-2\|(\mathbf{I} - \mathbf{P}_j)\mathbf{X}\beta\|^{-1}\epsilon'(\mathbf{I} - \mathbf{P}_j)\mathbf{X}\beta \\
&\ge \|(\mathbf{I} - \mathbf{P}_j)\mathbf{X}\beta\| - (3/2)h_n(k_0)\sigma^2\|(\mathbf{I} - \mathbf{P}_j)\mathbf{X}\beta\|^{-1}\} + o(1) \\
&\le \sum_{j=0}^{k_0-1} P\{-2\|(\mathbf{I} - \mathbf{P}_j)\mathbf{X}\beta\|^{-1}\epsilon'(\mathbf{I} - \mathbf{P}_j)\mathbf{X}\beta \\
&\ge (n\delta)^{1/2} - (3/2)h_n(k_0)\sigma^2(n\delta)^{-1/2}\} + \sum_{j=0}^{k_0-1} P(\|(\mathbf{I} - \mathbf{P}_j)\mathbf{X}\beta\|^2 < n\delta) + o(1) \\
&= \sum_{j=0}^{k_0-1} R_{j1} + \sum_{j=0}^{k_0-1} R_{j2} + o(1)
\end{aligned}
$$

for any constant $\delta > 0$.

It remains to show that R_{j1} and R_{j2} both converge to zero. By the Markov inequality and condition (B4),

$$\begin{aligned}
R_{j1} &\le 4(n\delta)^{-1}[1-3\sigma^2 h_n(k_0)/(2n\delta)]^{-2} \\
&\qquad E\{\|(\mathbf{I}-\mathbf{P}_j)\mathbf{X}\beta\|^{-2}\epsilon(\mathbf{I}-\mathbf{P}_j)\mathbf{X}\beta\beta'\mathbf{X}(\mathbf{I}-\mathbf{P}_j)\epsilon\} \\
&= 4\sigma^2(n\delta)^{-1}[1-3\sigma^2 h_n(k_0)/(2n\delta)]^{-2} \\
&\qquad E\{\|(\mathbf{I}-\mathbf{P}_j)\mathbf{X}\beta\|^{-2}\mathrm{tr}(\mathbf{I}-\mathbf{P}_j)\mathbf{X}\beta\beta'\mathbf{X}'(\mathbf{I}-\mathbf{P}_j)\} \\
&= 4\sigma^2(n\delta)^{-1}[1-3\sigma^2 h_n(k_0)/(2n\delta)]^{-2} \\
&\to 0.
\end{aligned}$$

To handle R_{j2}, note that k_0 and j are independent of n. Thus

$$\begin{aligned}
&n^{-1}\|(\mathbf{I}-\mathbf{P}_j)\mathbf{X}\beta\|^2 = n^{-1}\|(\mathbf{I}-\mathbf{P}_j)\mathbf{X}_{k_0}\beta_{k_0}\|^2 \\
&= \beta'_{k_0}\{(n^{-1}\mathbf{X}'_{k_0}\mathbf{X}_{k_0}) - n^{-1}\mathbf{X}'_{k_0}\mathbf{X}_j[n(\mathbf{X}'_j\mathbf{X}_j)^-]n^{-1}\mathbf{X}'_j\mathbf{X}_{k_0}\}\beta_{k_0} \\
&= \beta'_{k_0}\Xi_j\beta_{k_0} + o_P(1),
\end{aligned}$$

where, for $j < k_0$,

$$\Xi_j = E(\mathbf{x}_{k_0}\mathbf{x}'_{k_0}) - E(\mathbf{x}_{k_0}\mathbf{x}'_j)[E(\mathbf{x}_j\mathbf{x}'_j)]^{-1}E(\mathbf{x}_j\mathbf{x}'_{k_0}).$$

Taking $\delta = (1/2)\min_{j<k_0}\beta'_{k_0}\Xi_j\beta_{k_0}$, which is necessarily positive (see the proof of (3)), leads to $P(\|(\mathbf{I}-\mathbf{P}_j)\mathbf{X}\beta\|^2 \ge n\delta) \to 1$ for $j < k_0$. This implies $R_{j2} = o(1)$ and hence completes the proof.

Proof of Theorem 2.

It suffices to prove (9). Let ξ denote the maximum eigenvalue of $\mathbf{A}'\mathbf{A}$ and let $\|\mathbf{A}\| = \xi^{1/2}$ denote the matrix 2-norm of $\mathbf{A}$. Further, let $\mathbf{e}_i$ be the M_n-dimensional unit vector with its ith component equal to 1. By Lemma 1 of Mammen (1993) and condition (C2), $\|\mathbf{B}\| = o_P(1)$, where

$$\mathbf{B} = [E(\mathbf{x}_1\mathbf{x}'_1)]^{-1/2}(n^{-1}\mathbf{X}'\mathbf{X})[E(\mathbf{x}_1\mathbf{x}'_1)]^{-1/2} - \mathbf{I}.$$

Therefore by Theorem 10.3.1 of Campbell and Meyer (1979), with probability tending to one, $\mathbf{B}+\mathbf{I}$ and consequently $\mathbf{X}'\mathbf{X}$ are invertible. Without loss of generality we therefore assume that $(\mathbf{X}'\mathbf{X})^{-1}$ exists. Write $T_i = \hat\sigma^{-1}\hat\beta_i[\mathbf{e}'_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{e}_i]^{-1/2}$ and $\hat\beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon$.

If $i \notin \Gamma$, then $\beta_i = 0$ and $\hat\beta_i = \mathbf{e}'_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon$. By first conditioning on $\mathbf{X}$, we see that $E(\hat\sigma T_i) = 0$ and

$$\begin{aligned}
E(\hat\sigma^2 T_i^2) &= E\{[\mathbf{e}'_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{e}_i]^{-1}E[\epsilon'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{e}_i\mathbf{e}'_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon|\mathbf{X}]\} \\
&= \sigma^2. \qquad (17)
\end{aligned}$$

Therefore by condition (C2),

$$P[\max_{i\notin\Gamma}\hat{\sigma}|T_i| \geq n^{\frac{1}{2(1+\eta)}}] \leq \sum_{i\notin\Gamma} P(\hat{\sigma}|T_i| \geq n^{\frac{1}{2(1+\eta)}})$$
$$\leq \sum_{i\notin\Gamma} E(\hat{\sigma}^2 T_i^2) n^{-1/(1+\eta)} \leq \sigma^2 M_n n^{-1/(1+\eta)} = o(1). \qquad (18)$$

On the other hand, if $i \in \Gamma$,

$$\begin{aligned}\hat{\sigma}T_i &= \beta_i\{\mathbf{e}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{e}_i\}^{-1/2} + \mathbf{e}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon\{\mathbf{e}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{e}_i\}^{-1/2} \\ &= \beta_i\{\mathbf{e}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{e}_i\}^{-1/2} + O_P(1), \qquad (19)\end{aligned}$$

by (17). Since $\|\mathbf{B}\| = o_P(1)$, we have

$$\begin{aligned}\mathbf{e}_i'(n^{-1}\mathbf{X}'\mathbf{X})^{-1}\mathbf{e}_i &= \mathbf{e}_i'\{E(\mathbf{x}_1\mathbf{x}_1') - [E(\mathbf{x}_1\mathbf{x}_1') - n^{-1}\mathbf{X}'\mathbf{X}]\}^{-1}\mathbf{e}_i \\ &= \mathbf{e}_i'[E(\mathbf{x}_1\mathbf{x}_1')]^{-1/2}(\mathbf{I}+\mathbf{B})^{-1}[E(\mathbf{x}_1\mathbf{x}_1')]^{-1/2}\mathbf{e}_i \\ &= \mathbf{e}_i'[E(\mathbf{x}_1\mathbf{x}_1')]^{-1/2}(\mathbf{I}+\sum_{j\geq 1}(-1)^j\mathbf{B}^j)[E(\mathbf{x}_1\mathbf{x}_1')]^{-1/2}\mathbf{e}_i \\ &= \mathbf{e}_i'[E(\mathbf{x}_1\mathbf{x}_1')]^{-1}\mathbf{e}_i + R_i,\end{aligned}$$

where (c.f. Mammen (1993) or Golub and Van Loan (1989), Section 2.3)

$$\begin{aligned}|R_i| &= \Big|\mathbf{e}_i'[E(\mathbf{x}_1\mathbf{x}_1')]^{-1/2}\sum_{j\geq 1}(-1)^j\mathbf{B}^j[E(\mathbf{x}_1\mathbf{x}_1')]^{-1/2}\mathbf{e}_i\Big| \\ &\leq \|[E(\mathbf{x}_1\mathbf{x}_1')]^{-1/2}\|^2 \cdot \|\sum_{j\geq 1}\mathbf{B}^j\| \\ &\leq \|[E(\mathbf{x}_1\mathbf{x}_1')]^{-1/2}\|^2 \cdot \|\mathbf{B}\|(1-\|\mathbf{B}\|)^{-1} = o_P(1)\end{aligned}$$

by (C1). Equality (19) implies that, for $i \in \Gamma$,

$$\begin{aligned}\hat{\sigma}n^{-1/2}T_i &= \beta_i\{\mathbf{e}_i'[E(\mathbf{x}_1\mathbf{x}_1')]^{-1}\mathbf{e}_i + o_P(1)\}^{-1/2} + o_P(1) \\ &= \beta_i\{\mathbf{e}_i'[E(\mathbf{x}_1\mathbf{x}_1')]^{-1}\mathbf{e}_i\}^{-1/2} + o_P(1).\end{aligned}$$

Note that the first term on the right side of the last equality is bounded away from zero because by condition (C1), $\mathbf{e}_i'[E(\mathbf{x}_1\mathbf{x}_1')]^{-1}\mathbf{e}_i \leq \kappa^{-1} < \infty$. Since Γ is independent of n, it follows that

$$P(\min_{i\in\Gamma}\hat{\sigma}|T_i| \geq n^{\frac{1}{2(1+\eta)}}) \to 1, \quad \eta > 0.$$

This and (18) yield (9).

Proof of Theorem 3.

The method used to prove Theorem 1 still works here with some technical modifications to accommodate dependency among observations. We follow the two steps there with the same definition of $\tilde{k}_0$.

1. $\tilde{k}_0 \to_P k_0$. As in the proof of Theorem 1, (c.f. the inequality before (15)), we have

$$1-P(\tilde{k}_0 = k_0) \le P[\epsilon'\mathbf{P}_{M_n}\epsilon > \{h_n(k_0+1)-h_n(k_0)\}\sigma^2/2]+P(|\hat{\sigma}^2-\sigma^2| \ge \sigma^2/2),$$

which converges to zero if we can show that (c.f. (14))

$$\epsilon'\mathbf{P}_{M_n}\epsilon = O_p(M_n). \tag{20}$$

Write the autocovariance matrix $E(\mathbf{x}_1\mathbf{x}_1') = \Sigma_{M_n} = \gamma_0(\sigma_{ij})_{1\le i,j\le M_n}$ where $\gamma_0 = var(y_1)$ and $\sigma_{ij} = \rho_{|i-j|}$ is the autocorrelation defined in (12). It is known that Σ_{M_n} is nonsingular for all $1 \le M_n \le n$ (Hannan (1973)). Let $\mathbf{B} = \Sigma_{M_n}^{-1/2}(n^{-1}\mathbf{X}'\mathbf{X})\Sigma_{M_n}^{-1/2} - \mathbf{I}$. From Lemma 1 below, $\|\mathbf{B}\| = o_P(1)$.
Now (20) becomes

$$\begin{aligned}\epsilon'\mathbf{P}_{M_n}\epsilon &= n^{-1}(\Sigma_{M_n}^{-1/2}\mathbf{X}'\epsilon)'(\mathbf{I}+\mathbf{B})^{-1}(\Sigma_{M_n}^{-1/2}\mathbf{X}'\epsilon)\\ &= n^{-1}(\Sigma_{M_n}^{-1/2}\mathbf{X}'\epsilon)'\{\mathbf{I}+\sum_{j\ge 1}(-\mathbf{B})^j\}(\Sigma_{M_n}^{-1/2}\mathbf{X}'\epsilon)\\ &= n^{-1}\|\Sigma_{M_n}^{-1/2}\mathbf{X}'\epsilon\|^2 + R,\end{aligned}$$

where (c.f. the proof of Theorem 2)

$$|R| \le n^{-1}\|\Sigma_{M_n}^{-1/2}\mathbf{X}'\epsilon\|^2\|\mathbf{B}\|(1-\|\mathbf{B}\|)^{-1} = n^{-1}\|\Sigma_{M_n}^{-1/2}\mathbf{X}'\epsilon\|^2 \cdot o_P(1).$$

It thus suffices to establish that

$$n^{-1}\|\Sigma_{M_n}^{-1/2}\mathbf{X}'\epsilon\|^2 = O_P(M_n). \tag{21}$$

Recall that ϵ_i is independent of $\{y_{i-k},\ k \ge 1\}$ and hence of $\mathbf{x}_i' = (y_{i-1},\ldots, y_{i-M_n})$ for a stationary AR(k_0) process, and that each y_j can be written as a linear combination of $\{\epsilon_t,\ t \le j\}$ (Box and Jenkins (1976)). Therefore

$$E(\mathbf{X}'\epsilon\epsilon'\mathbf{X}) = E\Big\{\sum_{i=1}^{n}\mathbf{x}_i\epsilon_i\sum_{j=1}^{n}\epsilon_j\mathbf{x}_j'\Big\} = \sum_{i=1}^{n}E(\epsilon_i^2\mathbf{x}_i\mathbf{x}_i') = n\sigma^2\Sigma_{M_n},$$

where each $E(\mathbf{x}_i\epsilon_i\epsilon_j\mathbf{x}_j') = 0$ for $i \ne j$ by first conditioning on $\{\epsilon_t,\ t \le \max(i,j)-1\}$. The expectation of the left hand side of (21) is thus

$$n^{-1}E\{\epsilon'\mathbf{X}\Sigma_{M_n}^{-1}\mathbf{X}'\epsilon\} = n^{-1}\mathrm{tr}\{\Sigma_{M_n}^{-1}E[\mathbf{X}'\epsilon\epsilon'\mathbf{X}]\} = \sigma^2 M_n.$$

This proves (21).

2. $\hat{k}_0 - \tilde{k}_0 = o_P(1)$. As in Step 2 of the proof of Theorem 1,

$$P(|\hat{k}_0 - \tilde{k}_0| \neq 0) \leq \sum_{j=0}^{k_0-1} P\{n^{-1}\mathrm{RSS}_j - n^{-1}\mathrm{RSS}_{k_0} \leq n^{-1}[h_n(k_0) - h_n(j)]\hat{\sigma}^2\}. \tag{22}$$

It follows from Potscher (1989), proof of Lemma 3.3 that for each $j < k_0$,

$$\liminf_n \{n^{-1}\mathrm{RSS}_j - n^{-1}\mathrm{RSS}_{k_0}\} > 0, \text{ a.s.}$$

Condition (B4) and the fact that $\hat{\sigma}^2 \to_P \sigma^2$ imply that (22) converges to zero.

Lemma 1. *Suppose that the conditions of Theorem* 3 *hold. Then*

1. *The minimum eigenvalue* λ_{M_n} *of* Σ_{M_n} *satisfies* $\lambda_{M_n} \geq \lambda > 0$ *for some constant* λ *independent of* n.
2. $\|\mathbf{B}\| = o_P(1)$.

Proof. The proof makes use of Toeplitz forms in Grenander and Szego (1984), Chapters 5 and 10 and is available from the first author.

References

An, H. and Gu, L. (1985). On the selection of regression variables. *ACTA Mathematicae Applicatae Sinica* **2**, 27-36.

Bickel, P. J. and Freedman, D. A. (1983). Bootstrapping regression models with many parameters. In *A Festschrift for Erich L. Lehmann* (Edited by P. J. Bickel et al.), 28-48. Wadsworth, Belmont, CA.

Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis, Forecasting and Control*. Holden-Day, San Francisco.

Breiman, L. and Freedman, D. (1983). How many variables should be entered in a regression equation? *J. Amer. Statist. Assoc.* **78**, 131-136.

Breiman, L. and Spector, P. (1992). Submodel selection and evaluation in regression. The X-random case. *Internat. Statist. Rev.* **60**, 291-319.

Burman, P. (1989). A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika* **76**, 503-514.

Campbell, S. L. and Meyer, C. D. (1979). *Generalized Inverses of Linear Transformations*. Dover, New York.

Choi, B. S. (1992). *ARMA Model Identification*. Springer, New York.

Daniel, C. and Wood, F. S. (1980). *Fitting Equations to Data: Computer Analysis of Multifactor Data*. Wiley, New York.

Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. *J. Amer. Statist. Assoc.* **78**, 316-331.

Golub, G. H. and Van Loan, C. F. (1989). *Matrix Computations*. Johns Hopkins Press, Baltimore.

Grenander, U. and Szego, G. (1984). *Toeplitz Forms and Their Applications*. Chelsea Publishing Co., New York.

Hannan, E. J. (1973). The asymptotic theory of linear time-series models. *J. Appl. Probab.* **10**, 130-145.

Hannan, E. J. (1980). The estimation of the order of an ARMA process. *Ann. Statist.* **8**, 1071-1081.

Hannan, E. J. and Quinn, B. G. (1979). The determination of the order of an autoregression. *J. Roy. Statist. Soc. Ser. B* **41**, 190-195.

Hocking, R. R. (1976). The analysis and selection of variables in linear regression. *Biometrics* **32**, 1-49.

Huber, P. J. (1981). *Robust Statistics*. Wiley, New York.

Linhart, H. and Zucchini, W. (1986). *Model Selection*. Wiley, New York.

Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *Ann. Statist.* **21**, 255-285.

Miller, A. J. (1990). *Subset Selection in Regression*. Chapman and Hall, London.

Potscher, B. M. (1989). Model selection under nonstationarity: Autoregressive models and stochastic linear regression models. *Ann. Statist.* **17**, 1257-1274.

Rissanen, J. (1986). Stochastic complexity and modeling. *Ann. Statist.* **14**, 1080-1100.

Seber, G. A. F. (1984). *Multivariate Observations*. Wiley, New York.

Shao, J. (1993). Linear model selection by cross-validation. *J. Amer. Statist. Assoc.* **88**, 486-494.

Shibata, R. (1984). Approximate efficiency of a selection procedure for the number of regression variables. *Biometrika* **71**, 43-49.

Thompson, M. L. (1978). Selection of variables in multiple regression, Parts I and II. *Internat. Statist. Rev.* **46**, 1-19, 129-146.

Wei, C. Z. (1992). On predictive least squares principles. *Ann. Statist.* **20**, 1-42.

Zhang, P. (1992). On the distributional properties of model selection criteria. *J. Amer. Statist. Assoc.* **87**, 732-737.

Zhang, P. (1993). Model selection via multifold cross validation. *Ann. Statist.* **21**, 299-313.

Zheng, X. and Loh, W.-Y. (1994). Consistent bootstrap variable selection in linear models. Manuscript.

Zheng, X. and Loh, W.-Y. (1995). Consistent variable selection in linear models. *J. Amer. Statist. Assoc.* **90**, 151-156.

Department of Mathematics and Statistics, Utah State University, Logan, UT 84322-3900, U.S.A.

Department of Statistics, University of Wisconsin, Madison, WI 53706. U.S.A.

(Received December 1994; accepted October 1996)